DS 325 Applied Data Science – Data Storytelling Guide

All data can tell a story. As a data scientist, one of your responsibilities is to be that story teller. Throughout the semester, you will be asked to tell several data stories. A good place to start for ideas is the subreddit called DataIsBeautiful (www.reddit.com/r/dataisbeautiful), though you are welcome to use any data product you find, provided that you can address the questions below.

As part of the 4th hour activity for this course you will be required to author several Moodle blog posts along with a separate submission of answers to the questions below.

Guidelines for your Data Storytelling

Beginning with an article (or a post, or a plot, or a figure, or an infographic) that incorporates the analytical results from some data set (at least in part), you will be asked to address some specific questions:

For the blog post:

0. Summarize the information in the post/plot/figure/infographic that you chose. Your summary should be non-technical and meant for a general audience. You should include whatever information you think is necessary for someone to understand a) what the data is showing and b) why that's interesting. Your summary should be written as a short blog post to the Moodle forum along with the attached plot/figure/infographic you are discussing.

For your report to me:

1. Include your summary blog post. Research the (original) dataset yourself (or some representative subset of the data). From where does the dataset originate? Be sure to cite the data location (where you were able to retrieve it), the data author(s), and the date it was collected.

2. Provide a high-level summary of the dataset. In your summary you should include discussion of the size and dimensions of the dataset and how the data was obtained.

3. What specific features in the data were used to produce the article/post/plot/figure/infographic that you chose?

4. In looking over the data yourself, identify any scaling or transformations that were done between the raw data and the original post.

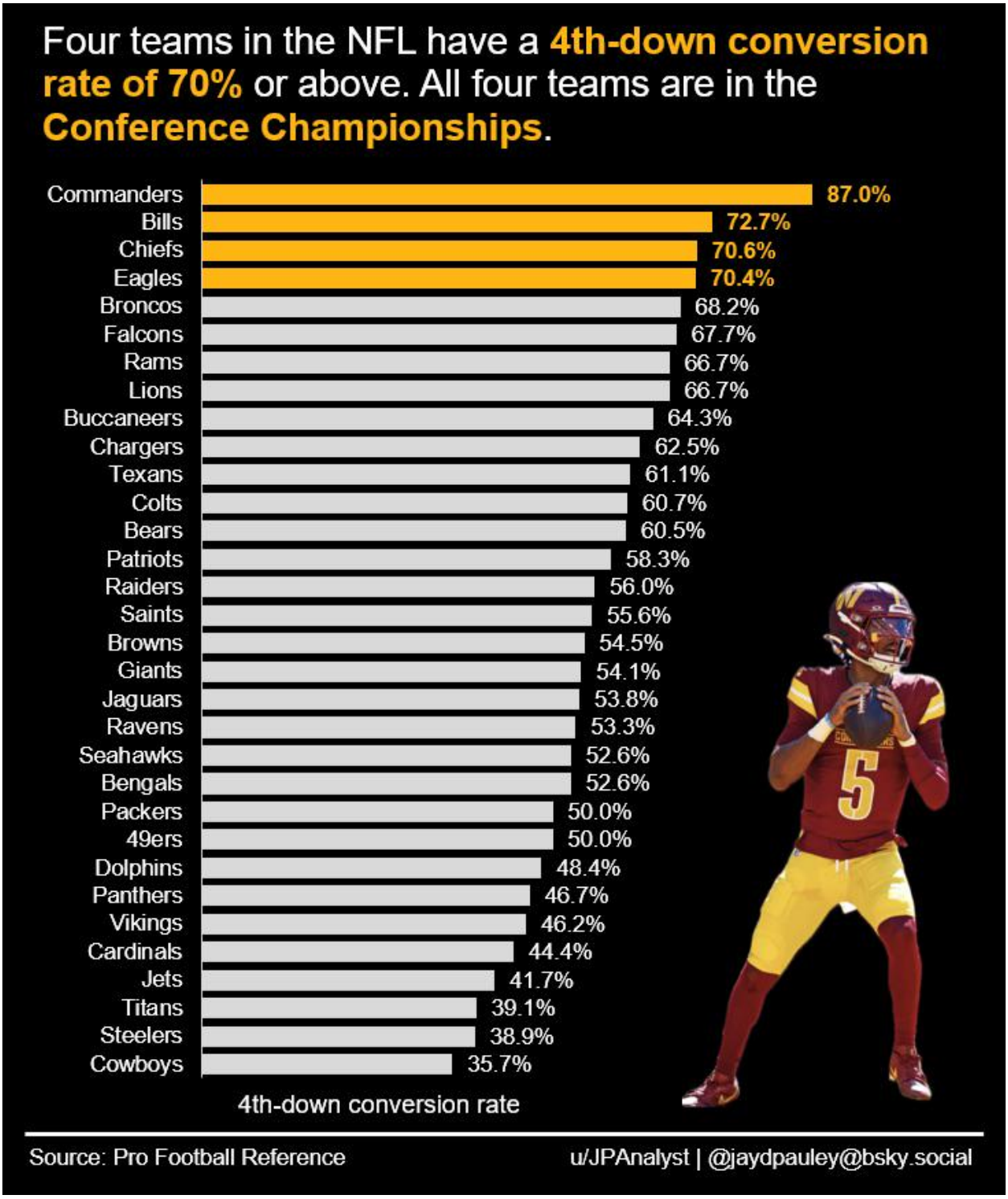
5. What story was the person who used this dataset trying to tell? Be specific. What was the purpose of creating the article (or post, or plot, or figure, or infographic)? If the purpose is not evident, then tell the story that you see in the data.

6. What are 1-2 additional stories that one could try and tell using the same data set? Describe what, if any, additional preparation or analysis would be required in order to tell these stories.

7. Are there additional questions you have about how the data was prepared or what further could be looked into? What are some improvements to the presentation you would suggest?

DS 325 Applied Data Science – Data Storytelling Example

As an example, I've done this for a recent /r/dataisbeautiful post: https://www.reddit.com/r/dataisbeautiful/comments/1i5qviw/oc_four_teams_in_the_nfl_have_a_4thdown/ submitted on Jan 20, 2025 which included the figure:



0. Summarize the information in the post/plot/figure/infographic that you chose.
Summary:

This figure contains stacked horizontal bar plots showing the percentage of successful attempts at 4th-down conversions for each NFL team. A 4th-down conversion attempt represents the final opportunity for the offense to either score or achieve a new set of downs. Whether successful or not, these conversion attempts are often critical decision-making turning points in football games. In this plot the 4 teams with the highest percentage success on 4th-down conversion attempts are highlighted in gold. It is noteworthy that the teams most successful in this metric are also the teams playing in next week's conference championship games, suggesting that this particular metric may be correlated with each team's overall success. There is also an image of the Washington Commanders quarterback Jayden Daniels, whose teams has the highest success rate at 4th down conversions.

1. Locate the (original) dataset yourself (or some representative subset of the data). From where does the dataset originate? Be sure to cite the data location, how you obtained that location, how you access the data, the data author(s), and the date the data was collected.

This post was published by reddit user "JPAnalyst" who appears to produce data products like for a living. As any good infographic should, this one cites the source data on the plot itself as: "Pro Football Reference"

A quick Google search of this phrase takes me to this site:
https://www.pro-football-reference.com
which appears to be an aggregate of many different NFL team and player statistics.

From there, I do a bit of navigating to try and find team statistics for the current year, which takes me here:
https://www.pro-football-reference.com/years/2024/#all_team_conversions

2. Provide a high-level summary of the dataset. In your summary you should include a discussion of the size and dimensions of the dataset and how the data was obtained.

The data contained in this table is, for each NFL team: # of games, 3rd down attempts, 3rd down conversions, 3rd down % success, 4th down attempts, 4th down conversions, **4th down % success**, red zone attempts, red zone TDs, red zone % success. I have highlighted the relevant data feature that appears to have been used in the plot.

The table these values are stored in may be automatically sorted in order to check against the original dataisbeautiful post:

Conversions

Share & Export ▼ Glossary Toggle Per-Game Stats

			Downs						Red Zone		
Rk	Tm	G	3DAtt	3DConv	3D%	4DAtt	4DConv	4D% ▼	RZAtt	RZTD	RZPct
1	Washington Commanders	17	217	99	45.6%	23	20	87.0%	71	45	63.4%
2	Buffalo Bills	17	202	89	44.1%	22	16	72.7%	67	48	71.6%
3	Kansas City Chiefs	17	229	111	48.5%	17	12	70.6%	65	35	53.8%
4	Philadelphia Eagles	17	235	98	41.7%	27	19	70.4%	68	39	57.4%

3. What specific features in the data were used to produce the article/post/plot/figure/infographic that you chose?

Specifically, the feature entitled 4D% was the one used to produce the plot shown. It is actually a result of combining the previous 2 columns which are 4DAtt and 4DConv, measuring the number of 4th down attempts and successes, respectively. The feature 4D% is computed as the quotient of 4DConv/4DAtt.

4. In looking over the data yourself, identify any scaling or transformations that were done between the raw data and the original post.

It does not appear there was any scaling or transformation of the data from the original data table into the plot that is shown. In making the visual representation, stacked horizontal bar charts are shown with the team having the largest value on top and the smallest at the bottom.

5. What story was the person who used this dataset trying to tell? Be specific. What was the purpose of creating the article (or post, or plot, or figure, or infographic)? If the purpose is not evident, then tell the story that you see in the data.

The purpose of this post appears to be drawing a correlation between a team's success at 4th down conversions and their overall success, since the top 4 teams are the only remaining teams with a chance to win this year's Super Bowl.

6. What are 1 (if you gave an example in #5 already) or 2 (if you didn't give an example in #5) additional stories that one could try and tell using the same data set? Describe what, if any, additional preparation or analysis would be required in order to tell these stories.

Rather than focusing on the teams at the top, one could also ask whether being at the bottom of this pack is correlated with lower overall success.

7. Are there additional questions you have about how the data was prepared or what further could be looked into? What are some improvements to the presentation you would suggest?

One reason noting this correlation could be important is in a predictive sense. However, in order to use this statistic as predictive, one would have to exam it historically and see if the trend this year were repeated, even in part, at some point in the past. Perhaps looking at a timeline of the top 4 teams in this statistic each year, and whether they played in their respective conference championships would be a good place to start.